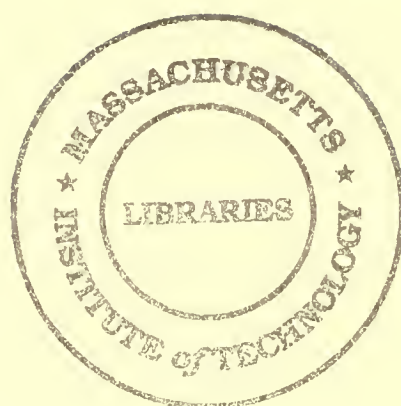


MIT LIBRARIES



3 9080 02239 0238

BACEMENT



WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

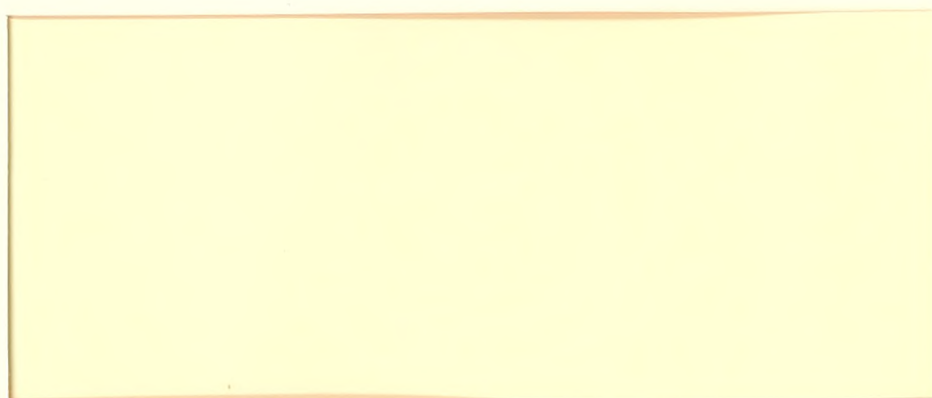
On the Convergence of Multiclass Queueing
Networks in Heavy Traffic

J.G. Dai
and
Viên Nguyen

WP# 3489-92-MSA

October, 1992

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139



On the Convergence of Multiclass Queueing
Networks in Heavy Traffic

J.G. Dai
and
Viên Nguyen

WP# 3489-92-MSA

October, 1992

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

NOV 14 2000

LIBRARIES

On the Convergence of Multiclass Queueing Networks in Heavy Traffic

J. G. Dai¹ and Viêt Nguyen

Georgia Institute of Technology and Massachusetts Institute of Technology

Abstract

The subject of this paper is the heavy traffic behavior of a general class of queueing networks with first-in-first-out (FIFO) service discipline. For special cases which require various assumptions on the network structure, several authors have proved heavy traffic limit theorems to justify the approximation of queueing networks by reflected Brownian motions (RBM's). Based on these theorems, some have conjectured that the Brownian approximation may in fact be valid for a more general class of queueing networks.

In this paper, we prove that the Brownian approximation does *not* hold for such a general class of networks. Our findings suggest that studying Brownian models of non-FIFO queueing networks may perhaps be more fruitful.

Contents

1. Introduction
2. Preliminaries
3. Conjecture and the Main Theorem
4. A Pseudo Heavy Traffic Limit Theorem
5. Proof of Theorem 3.1
6. Concluding Remarks and Open Problems

Keywords: Multiclass queueing network, heavy traffic, diffusion approximation, reflecting Brownian motion, performance analysis.

AMS 1991 subject classifications: Primary 60K25, 60F17, 60G17; Secondary 60J15, 90B22, 68M20.

October, 1992

¹Research supported in part by two grants from Texas Instruments Corporation and by NSF grant DMS 92-09586

1 Introduction

The past few years have witnessed a surge in research activities dealing with Brownian approximations of queueing networks [3, 5, 7, 10, 15, 16]. This line of research suggests that certain processes associated with a queueing network can be approximated by reflected Brownian motions (RBM's), and these approximations are justified by so called "heavy traffic limit theorems." Although such limit theorems have been proved only for special cases which require various assumptions on the network structure, some authors have proposed that the Brownian approximations in fact may be used for a more general class of networks that operate under the first-in-first-out service discipline [15, 16].

In this paper, we prove that the Brownian approximation is *not* valid for such a general class of networks. We do so by proving a "pseudo" heavy traffic limit theorem, which states that if the process associated with the queueing network converges to a continuous limit, then that limit must be the RBM specified by the Brownian approximation. We then present a queueing network example developed by Dai and Wang [11] for which the specified Brownian approximation is not well defined. Our findings suggest that it may be fruitful to consider a more general class of approximating processes. Furthermore, our findings signal that other service disciplines may yield more tractable structures.

We consider a network composed of d single server stations, which we index by $j = 1, \dots, d$. The network is populated by c classes of customers, and each class k has its own exogenous renewal arrival process $E_k = \{E_k(t), t \geq 0\}$ (possibly null), where $E_k(t)$ is the number of class k customers who arrive at the network by time t . For each customer class $k = 1, \dots, c$, it is assumed that $E_k(0) = 0$ and customer inter-arrival times have mean $1/\alpha_k$ with squared coefficient of variation (SCV) $c_{a,k}^2$. (The SCV of a random variable is defined as its variance divided by the square of its mean.) We denote by E the c -dimensional process with components E_1, \dots, E_c . (All vectors are envisioned as column vectors.) We assume that arrival processes E_1, \dots, E_c are independent and $\alpha_k > 0$ for at least one k . For each k , α_k is interpreted as the *long-run average arrival rate* for class k customers. These customers require service at station $s(k)$, and their service times are independent and identically distributed (i.i.d.) with mean m_k and SCV $c_{s,k}^2$. The service time sequences for the various customer classes are independent of one another and are also independent of the arrival processes. Upon completion of service at station $s(k)$, a class k customer becomes a customer of class l with probability P_{kl} and exits the network with probability $1 - \sum_l P_{kl}$, independent of all previous history. The transition matrix $P = (P_{kl})$ is taken to be transient, which simply means that all customers eventually leave the network. Hence the networks we are considering are open queueing networks. We assume that the waiting buffer at each station has infinite capacity, and that customers are served at each station on a first-in-first-out (FIFO) basis. Hereafter, we will refer to such a network as a *multiclass open queueing network*.

Such a description of a multiclass network is now quite standard, as in Harrison and Nguyen [15, 16]. (The class of queueing networks described here is in fact an important special case of the setup in [15, 16].) Figure 1 shows an example of such a network, which Dai and Wang [11] have studied. Customers arrive at station 1 according to a Poisson process with rate $\alpha_1 = 1$. Each customer follows a deterministic route whose sequence of station visitations is 1, 1, 2, 2, 1, after which the customer departs from the network. Hence, each customer makes

5 stops before exiting the network, and we designate those customers in their k^{th} stop as class k customers.

In his pioneering paper on queueing networks, Jackson [22] assumed that customers visiting or occupying any given station are essentially indistinguishable from one another, and that a customer completing service at station i will move next to station j with some fixed probability P_{ij} , independent of all previous history. Thus in Jackson's networks, each station serves a single customer class, hence these networks have been called *single-class* networks. Jackson's model was extended by Baskett *et al.* [1] and Kelly [24] to networks populated by multiple types of customers, each type following a deterministic route. The routing mechanism described in this paper subsumes those considered in [1, 24]. Readers are referred to Harrison [14] and Harrison and Nguyen [15, 16] for further discussion.

For each $j = 1, \dots, d$ and each $t \geq 0$, let $W_j(t)$ denote the sum of the impending service times for all customers who are queued at station j at time t , plus the remaining service time for any customer who may be in service there at time t . If a new customer arrives to station j at time t , that customer must wait $W_j(t)$ time units before gaining access to the server, so one can also describe $W_j(t)$ as the virtual waiting time process for station j . Set $W(t)$ to be the d -vector with components $W_1(t), \dots, W_d(t)$. Define the process $W_j = \{W_j(t), t \geq 0\}$ and let W be the corresponding d -dimensional workload process defined in the natural way.

Intuitively, when the system is heavily loaded, the workload $W(t)$ at time t will be large for large t . Let ρ_j be the *traffic intensity* at station j (this term will be defined in Section 2). The quantity ρ_j can also be interpreted as the long-run fraction of time that server j is busy. As an example, because the arrival rate is set to be 1, the traffic intensities for the network pictured in Figure 1 are given by $\rho_1 = m_1 + m_2 + m_5$ and $\rho_2 = m_3 + m_4$. To facilitate our explanation of the underlying concepts, let us for the moment assume that

$$(1.1) \quad \rho_j = 1, \quad j = 1, \dots, d.$$

Condition (1.1) is a special form of the *heavy traffic condition* as described in Section 2. For fixed $t \geq 0$, we are interested in how fast $W(nt)$ goes to infinity as $n \rightarrow \infty$. It has been widely believed that

$$(1.2) \quad \tilde{W}^n(t) \equiv \frac{1}{\sqrt{n}} W(nt) \Rightarrow W^*(t) \quad \text{as } n \rightarrow \infty,$$

where $W^* = \{W^*(t), t \geq 0\}$ is a d -dimensional semimartingale reflecting Brownian motion (SRBM) and the symbol " \Rightarrow " denotes weak convergence (the notion of weak convergence will be clarified in Section 2). Our main contribution in this paper is the proof that conjecture (1.2) does not hold in general.

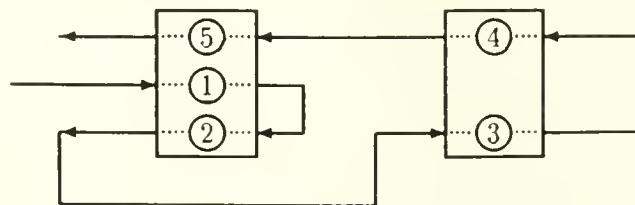


Figure 1: A two-station network with self-feedback

QNET Method for Performance Analysis

Exact analysis of a multiclass network models is limited essentially to networks whose inter-arrival times have exponential distributions and whose service times for *all* customers at each station have the *same exponential* distributions (cf. Baskett *et al.* [1]). Such networks are commonly referred to as BCMP or Kelly networks. For networks with other distributional forms, no exact formulae are available to facilitate the analysis.

In recent years, a promising new approximation scheme, known as the QNET method, has been proposed. The first step in a QNET analysis is to replace one's "exact" queueing model by an approximating Brownian model, see Harrison and Nguyen [15, 16]. The second step involves steady-state analysis of the approximating Brownian model; see Harrison and Williams [18], Dai and Harrison [7] and Dai and Kurtz [8]. For a queueing network with d stations, this analysis requires that one determines the stationary distribution of a d -dimensional reflecting Brownian motion. Finally, summary statistics derived from that stationary distribution are used to obtain approximate steady-state performance measures for the original systems. Unlike previous approximations, the Brownian approximations culminate in estimates of complete distributions; readers can find examples of Brownian estimates for complete sojourn time distributions in Harrison and Nguyen [16]. Readers are referred to Harrison and Nguyen [15, 16], Dai and Harrison [6, 7] and Dai, Nguyen and Reiman [10] for more discussion on the QNET method, as well as its accuracy and efficiency. Those interested in the QNET software package can send an email message to dai@isye.gatech.edu.

Heavy Traffic Limit Theorems

In cases where (1.2) hold, the corresponding theorem is called a "heavy traffic limit theorem." There now exists a variety of heavy traffic limit theorems for networks with certain special structures. The first heavy traffic limit theorem for networks of queues is due to Iglehart and Whitt [20, 21], who worked with single-class queues in series. For single-class networks whose routing structure is similar to that of Jackson's networks, but whose inter-arrival times and service times may have general distributions, Reiman [26] proved that under the heavy traffic condition, the normalized queue length process converges weakly to a reflecting Brownian motion (RBM). The definition of an RBM was first presented in Harrison and Reiman [17]. Reiman's proof was later simplified by Johnson [23]. Reiman's result has been extended by Chen and Shanthikumar [5] to networks in which stations may have multiple servers. Peterson [25] proved an analogous heavy traffic limit theorem for multiclass feedforward networks. The term "feedforward" denotes a routing structure in which stations can be numbered so that customers always travel from lower numbered stations to higher numbered ones. Reiman [27] proved a heavy traffic limit theorem for a multiclass one station feedback queue. Dai and Kurtz [9] have greatly simplified Reiman's proof. A heavy traffic limit theorem for single-class closed networks was proved by Chen and Mandelbaum [3]. Strong approximations for single class networks were discussed in Glynn and Whitt [13] and Chen and Mandelbaum [4].

Until recently, it was believed that a heavy limit theorem should hold for multiclass open queueing networks of the type described in this section. Based on existing heavy traffic limit theorems, Harrison and Nguyen [15, 16] proposed Brownian models to approximate these

networks. Unfortunately, Dai and Wang [11] have found two-station and three-station networks for which Harrison and Nguyen's Brownian models fail to exist. (A more explicit interesting example showing no convergence was given by Whitt [30], in which he discovered chaotic behavior for certain multiclass open queueing networks.) Building on Dai and Wang's example, we prove in this paper the following general result: *There exist multiclass open queueing networks for which the scaled workload process $\tilde{W}^n(t)$ does not converge to any continuous limits.* A by-product of our result is that if the normalized workload process converges to a continuous limiting process, that process must be the reflected Brownian motion identified by Harrison and Nguyen.

Organization of the Paper

We introduce some additional notation and definitions in Section 2. In Section 3, we state the heavy traffic conjecture and the main theorem of this paper. We prove our theorem by way of a “pseudo heavy traffic limit theorem,” which we state and justify in Section 4. The proof of our main theorem is in Section 5. Finally, we discuss some open problems in Section 6.

2 Preliminaries

We now define several processes that will be used later in later sections. Let $\{\phi^k(1), \phi^k(2), \dots\}$ be a sequence of i.i.d. *routing vectors* for class k customers. The l^{th} component of the vector $\phi^k(i)$ equals 1 if the i^{th} class k customer next goes to class l , and all other components are zero. Also, define the c -dimensional cumulative sum processes

$$(2.1) \quad \Phi^k(r) = \phi^k(1) + \dots + \phi^k(r).$$

Finally, set $\mathcal{C}(j)$ to be the set of all customer classes k that receive service at station j , that is, $\mathcal{C}(j) = \{k : s(k) = j\}$. This set is called the *constituency* of server j in Harrison [14]. We require that $\mathcal{C}(j)$ be nonempty for each $j = 1, \dots, d$.

Next, set C to be the $d \times c$ incidence matrix with components

$$(2.2) \quad C_{jk} = \begin{cases} 1 & \text{if } k \in \mathcal{C}(j), \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $E_k(t)$ is the *external* arrival process for class k . Denote by $A_k(t)$ the *total* number of customer visits to class k by time t and by $D_k(t)$ the total number of customer departures from class k by time t . Letting $F_k(t)$ denote the number of visits to class k by time t that result from *internal* transitions, one has as a matter of definition

$$(2.3) \quad A_k(t) = E_k(t) + F_k(t) = E_k(t) + \sum_{i=1}^c \Phi_k^i(D_i(t)).$$

Let $\{v_k(1), v_k(2), \dots\}$ be a sequence of i.i.d. service times associated with class k customers, and let $V_k(r)$ be the cumulative sum process defined by

$$V_k(r) = v_k(1) + \dots + v_k(r).$$

We denote by $V(A(t))$ the c -dimensional process whose k^{th} component is given by $V_k(A_k(t))$, and we set

$$(2.4) \quad L(t) = CV(A(t)).$$

Note that $L = \{L(t), t \geq 0\}$ is a d -dimensional process; one interprets $L_j(t)$ as the amount of work for server j brought by all those customers who have arrived at station j by time t . The process L_j was referred to as the *immediate workload input* process for station j by Harrison and Nguyen [16].

Let $Y_j(t)$ be the amount of cumulative idleness experienced by server j up to time t and let $Y(t) = (Y_1(t), \dots, Y_d(t))'$ be the corresponding vector process. (Prime denotes transpose.) We can express the d -dimensional workload process $W = \{W(t), t \geq 0\}$ as follows:

$$(2.5) \quad W(t) = L(t) - te + Y(t),$$

where e is the d -dimensional vector of ones. Clearly, the idleness process $Y_j(\cdot)$ may increase only at times t such that $W_j(t) = 0$, hence

$$(2.6) \quad Y_j(t) = - \inf_{0 \leq s \leq t} \{L_j(s) - s\}, \quad j = 1, \dots, d.$$

To finish our description of the network model, let us define the fundamental matrix

$$(2.7) \quad Q = (I - P')^{-1} = (I + P + P^2 + \dots)'$$

The $(k, l)^{\text{th}}$ element of Q represents the expected number of visits to class k made by a customer who starts in class l . Thus, defining $\lambda = (\lambda_1, \dots, \lambda_c)'$ via

$$(2.8) \quad \lambda = Q\alpha,$$

one recognizes that λ_k is the long-run average number of customer visits to class k per unit time (resulting from external arrivals as well as internal transitions). The total traffic intensity at station j is then defined by

$$(2.9) \quad \rho_j = \sum_{k \in \mathcal{C}(j)} \lambda_k m_k.$$

Let ρ be the vector of traffic intensities at stations $1, \dots, d$. One can express the vector of traffic intensities in matrix form via

$$(2.10) \quad \rho = CM\lambda,$$

where M is the $c \times c$ diagonal matrix with diagonal elements m_1, \dots, m_c .

To rigorously state our convergence result, we need to introduce the path space $D^c[0, \infty)$, which is the space all functions $f : [0, \infty) \rightarrow \mathbb{R}^c$ which are right continuous on $[0, \infty)$ and have finite left limits on $(0, \infty)$. The path space $D^c[0, \infty)$ is endowed with the Skorohod topology, see Billingsley [2]. For a sequence $\{X^n\}$ of $D^c[0, \infty)$ -valued stochastic processes and $X \in D^c[0, \infty)$, we write $X^n(\cdot) \Rightarrow X(\cdot)$ if X^n converges to X in distribution.

For a function $f : [0, \infty) \rightarrow \mathbb{R}$ and $t \geq 0$, put

$$\|f\|_t \equiv \sup_{0 \leq s \leq t} |f(s)|,$$

and for a vector of functions $f = (f_1, \dots, f_k)' : [0, \infty) \rightarrow \mathbb{R}^k$ and $t \geq 0$, put

$$\|f\|_t = (\|f_1\|_t, \dots, \|f_k\|_t)'.$$

A sequence $\{f^n\}$ of functions $f^n : [0, \infty) \rightarrow \mathbb{R}^k$ is said to converge *uniformly on compact* (u.o.c.) sets to $f : [0, \infty) \rightarrow \mathbb{R}^k$ if for each $t \geq 0$, $\|f^n - f\|_t \rightarrow 0$ as $n \rightarrow \infty$. For a sequence $\{X^n\}$ of $D^c[0, \infty)$ -valued stochastic processes and $X \in D^c[0, \infty)$ defined on a probability space, we write $X^n(\cdot) \rightarrow X(\cdot)$ u.o.c. if almost surely, X^n converges to X uniformly on compact sets.

3 Conjecture and the Main Theorem

In order to rigorously state a heavy traffic limit theorem, we need to consider a “sequence of networks” indexed by n . Our setup here follows closely that of Harrison and Nguyen [16]. Let α^n and m^n be vectors of inter-arrival rates and service times, respectively, associated with the n^{th} network in the sequence. We may assume without loss of generality, however, that the routing matrix and the squared coefficients of variation for inter-arrival times and service times remain fixed across the sequence of networks. Define ρ^n to be the vector of traffic intensities for the n^{th} matrix similarly to (2.9). We are interested in the sequence of networks such that

$$(3.1) \quad \alpha^n \rightarrow \alpha, \quad m^n \rightarrow m > 0,$$

and

$$(3.2) \quad \sqrt{n}(\rho^n - e) \rightarrow \beta, \quad \text{as } n \rightarrow \infty,$$

where e , as before, is the vector of ones. Condition (3.2) requires that $\rho_j^n \rightarrow 1$ at an appropriate rate, and is known as the *heavy traffic condition*. As $n \rightarrow \infty$, we are interested in the limit of the normalized workload process \tilde{W}^n defined by

$$(3.3) \quad \tilde{W}^n(t) = \frac{1}{\sqrt{n}} W^n(nt), \quad t \geq 0.$$

Before we state the conjecture, we define more scaled processes. For each $t \geq 0$ and $n \geq 1$, set

$$\begin{aligned} \tilde{E}^n(t) &= \frac{1}{\sqrt{n}} (E^n(nt) - \alpha^n nt), \\ \tilde{V}^n(t) &= \frac{1}{\sqrt{n}} (V^n([nt]) - m^n nt), \\ \tilde{\Phi}_k^{i,n}(t) &= \frac{1}{\sqrt{n}} (\Phi_k^i([nt]) - P_{ik} nt), \quad i = 1, \dots, c, \quad k = 1, \dots, c, \end{aligned}$$

where $[x]$ is the integer part of x . (Again, note that the processes Φ^i do not change with n .) It follows from the classical Donsker theorem that as $n \rightarrow \infty$

$$(3.4) \quad \tilde{E}^n \Rightarrow \xi^a,$$

$$(3.5) \quad \tilde{V}^n \Rightarrow \xi^s,$$

$$(3.6) \quad \tilde{\Phi}^{i,n} \Rightarrow \xi^i, \quad i = 1, \dots, c.$$

where ξ^a , ξ^s and ξ^i ($i = 1, \dots, c$) are $(c + 2)$ independent c -dimensional zero-drift Brownian motions with covariance matrices Γ^a , Γ^s and Γ^i ($i = 1, \dots, c$), respectively. It is easily verified that $\Gamma^a = \text{diag}(\alpha_1 c_{a,1}^2, \dots, \alpha_c c_{a,c}^2)$, $\Gamma^s = \text{diag}(m_1^2 c_{s,1}^2, \dots, m_k^2 c_{s,c}^2)$ and Γ^i is a matrix defined by

$$\Gamma_{kl}^i = \begin{cases} P_{ik}(1 - P_{ik}) & \text{if } k = l \\ -P_{ik}P_{il} & \text{if } k \neq l. \end{cases}$$

Because Brownian motions are continuous and \tilde{E}^n , \tilde{V}^n , $\tilde{\Phi}^{i,n}$ ($i = 1, \dots, c$) are independent, we can and will assume by the Skorohod representation theorem that the convergence in (3.4)–(3.6) holds u.o.c.

Conjecture 1 *Under the heavy traffic conditions (3.1)–(3.2), the sequence of normalized workload processes \tilde{W}^n defined in (3.3) converges to a continuous process $W^* = \{W^*(t), t \geq 0\}$ uniformly on compact sets as $n \rightarrow \infty$. That is*

$$(3.7) \quad \frac{1}{\sqrt{n}} W^n(nt) \rightarrow W^*(t), \quad \text{u.o.c., as } n \rightarrow \infty.$$

Theorem 3.1 *There exist multiclass open queueing networks for which the scaled workload process \tilde{W}^n does not converge to any continuous limit. In particular, Conjecture 1 is false.*

The key to the proof of Theorem 3.1 is the “pseudo” heavy traffic result stated in Theorem 4.1 and proved in the next section together with the Dai-Wang example in [11]. We leave the proof of Theorem 3.1 to Section 5.

Corollary 3.1 *If the conclusion in conjecture 1 is changed to $\{\tilde{W}^n, n \geq 1\}$ being D -tight, the conjecture is still false.*

The definition of tightness is given in, for example, Section 3.2 of Ethier and Kurtz [12]. The proof of the corollary is given at the end of Section 5.

4 A Pseudo Heavy Traffic Limit Theorem

Set

$$G = CMQP'\Lambda C',$$

where $M = \text{diag}(m_1, \dots, m_c)$ and $\Lambda = \text{diag}(\lambda)$, and $\text{diag}(\lambda)$ is the diagonal matrix with diagonal elements $\lambda_1, \dots, \lambda_c$. Recall that $Y_j^n(t)$ is the cumulative idleness of server j by time t for the n^{th} system, and $Y^n(t)$ is the d -dimensional vector with components Y_1^n, \dots, Y_d^n . Set $\tilde{Y}^n(t) = n^{-1/2} Y^n(nt)$.

Theorem 4.1 *Assume Conjecture 1 is true, namely, that the convergence in (3.7) holds. Then the sequence of normalized idleness process \tilde{Y}^n converges to Y^* u.o.c. and the limiting processes*

(W^*, Y^*) must satisfy

$$(4.1) \quad (I + G)W^*(t) = C\xi^s(\lambda t) + CMQ \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) \right) + \beta t + Y^*(t).$$

$$(4.2) \quad W^*(t) \geq 0,$$

$$(4.3) \quad Y^*(0) = 0, \ Y^* \text{ is continuous and nondecreasing,}$$

$$(4.4) \quad Y_j^*(\cdot) \text{ increases only at times } t \text{ such that } W_j^*(t) = 0, \ j = 1, \dots, d.$$

Remark. Theorem 4.1 states that if (3.7) is true, then the Brownian model proposed by Harrison and Nguyen [15, 16] is the correct model. Harrison and Nguyen summarily referred to this result (namely, the pseudo proof) in Section 5 of [16]. We offer a complete proof in this paper and consequently use this theorem to prove Theorem 3.1 in Section 5.

The remainder of this section is devoted to proving Theorem 4.1. We begin by introducing some important notation. For $j = 1, \dots, d$ and $t \geq 0$, define $\tau_j^n(t)$ to be the arrival time to station j of the customer currently being serviced there if $W_j^n(t) > 0$, and to be t if $W_j^n(t) = 0$. Let $\tau^n(t)$ be the d -dimensional vector defined in the obvious manner. This definition of $\tau^n(t)$, which is slightly different from what was given in Peterson [25, page 103], enables us to give a concise proof of Lemma 4.2 below. With $\tau_j^n(t)$, one can verify that the number of class k customers to have departed from station $j = s(k)$ by time t is given by

$$(4.5) \quad D_k^n(t) = \begin{cases} A_k^n(\tau_j^n(t)) - 1, & \text{if server } j \text{ is currently serving a class } k \text{ customer,} \\ A_k^n(\tau_j^n(t)), & \text{otherwise.} \end{cases}$$

We will use $A^n(\tau^n(t))$ to denote the c -dimensional process whose k^{th} component is $A_k^n(\tau_{s(k)}^n(t))$. For each $t \geq 0$ and $n \geq 1$, define

$$\tilde{\tau}^n(t) = \frac{1}{n} \tau^n(nt), \quad \tilde{\tau}^n(t) = \frac{1}{\sqrt{n}}(nt - \tau^n(nt)),$$

and

$$\bar{A}^n(t) = \frac{1}{n} A^n(nt), \quad \tilde{A}^n(t) = \frac{1}{\sqrt{n}} (A^n(nt) - \lambda^n t).$$

The first two lemmas below hold in general without the assumption of convergence in (3.7).

Lemma 4.1 *For each sample path and each $t \geq 0$, there exists κ independent of n such that*

$$\|\bar{A}_k^n(\cdot)\|_t \leq \kappa, \quad k = 1, \dots, c, \quad n \geq 1.$$

Proof. Let $S_k^n = \{S_k^n(t), t \geq 0\}$ be the renewal process associated with class k service times. Let $T_k^n(t)$ be the cumulative time that server $s(k)$ has devoted to class k customers in t units of time. Then, the number of class k customers to have departed from station $s(k)$ by time t is $D_k^n(t) = S_k^n(T_k^n(t)) \leq S_k^n(t)$. Therefore, from (2.3), we have

$$A^n(t) = E^n(t) + \sum_{k=1}^c \Phi^k(S_k^n(T_k^n(t))) \leq E^n(t) + \sum_{k=1}^c \Phi^k(S_k^n(t)).$$

The lemma then follows from the functional strong law of large numbers for random walks and renewal processes. \square

From the definition of $\tau_j^n(t)$, it follows that

$$(4.6) \quad t = \tau_j^n(t) + W_j^n(\tau_j^n(t)) - \epsilon_j^n(t).$$

where $\epsilon_j^n(t)$ is 0 if $W_j^n(t) = 0$, and otherwise is equal to the remaining service time of the customer currently occupying server j . Define $\tilde{\tau}_j^n(t) = t - \tau_j^n(t)$, and note that

$$W_j^n(\tau_j^n(t)) - \epsilon_j^n(t) = \tilde{\tau}_j^n(t) \leq W_j^n(\tau_j^n(t)).$$

The next three results show that under the heavy traffic scaling, the processes $\tilde{\tau}^n(t)$ and $\tilde{W}^n(t)$ are close for large n . We begin with the following lemma, which proves that $\epsilon_j^n(t)$ is negligible under the heavy traffic scaling.

Lemma 4.2 *For $j = 1, \dots, d$,*

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \epsilon_j^n(nt) = 0, \quad \text{u.o.c. as } n \rightarrow \infty.$$

Proof. It follows from the definition of $\epsilon_j^n(t)$ that

$$0 \leq \epsilon_j^n(t) \leq \max_{k \in \mathcal{C}(j)} \max_{1 \leq i \leq A_k^n(t)} v_k^n(i),$$

where $\{v_k^n(1), v_k^n(2), \dots\}$ is the sequence of i.i.d. service times for class k customers. An application of Lemma 3.3 from Iglehart and Whitt [20] yields

$$\left\| \frac{1}{\sqrt{n}} \epsilon_j^n(n \cdot) \right\|_t \rightarrow 0.$$

□.

Lemma 4.3 *Suppose the convergence in (3.7) holds. Then*

$$\tilde{\tau}^n(t) \rightarrow e t \quad \text{u.o.c.,}$$

where e is the d -dimensional vector of ones.

Proof. Let $\bar{W}_j^n(t) = \frac{1}{n} W_j^n(nt)$. Then,

$$\bar{W}_j^n(\tilde{\tau}_j^n(t)) - \frac{1}{n} \epsilon_j^n(nt) = \frac{1}{n} \tilde{\tau}_j^n(nt) \leq \bar{W}_j^n(\tilde{\tau}_j^n(t)),$$

Because $\tilde{\tau}^n(s) \leq s$ for $s \geq 0$,

$$\left\| \frac{1}{n} \tilde{\tau}_j^n(n \cdot) \right\|_t \leq \frac{1}{\sqrt{n}} \left\| \frac{1}{\sqrt{n}} W_j^n(n \cdot) \right\|_t + \frac{1}{\sqrt{n}} \left\| \frac{1}{\sqrt{n}} \epsilon_j^n(n \cdot) \right\|_t.$$

With the assumption of (3.7) and Lemma 4.2, the lemma is proved. □

Lemma 4.4 *Suppose the convergence in (3.7) holds. Then,*

$$\tilde{\tau}^n(t) \rightarrow W^*(t) \quad \text{u.o.c. as } n \rightarrow \infty.$$

Proof. Because

$$W_j^n(\tau_j^n(t)) - \epsilon_j^n(t) = t - \tau_j^n(t) \leq W_j^n(\tau_j^n(t)),$$

we have

$$\tilde{W}_j^n(\bar{\tau}_j^n(t)) - \frac{1}{\sqrt{n}}\epsilon_j^n(nt) = \bar{\tau}_j^n(t) \leq \tilde{W}_j^n(\bar{\tau}_j^n(t)).$$

The lemma follows immediately from assumption (3.7) and Lemmas 4.2 and 4.3. \square

Lemma 4.5 *Suppose the convergence in (3.7) holds. Then*

$$\bar{A}^n(t) \rightarrow \lambda t \quad \text{u.o.c.}$$

Proof. It follows from (2.3) that

$$\bar{A}^n(t) = \bar{E}^n(t) + \sum_{k=1}^c \bar{\Phi}^{k,n}(\bar{D}_k^n(t)),$$

where $\bar{E}^n(t) = \frac{1}{n}E^n(nt)$, $\bar{D}^n(t) = \frac{1}{n}D^n(nt)$ and $\bar{\Phi}^{k,n}(t) = \frac{1}{n}\Phi^k([nt])$ for $k = 1, \dots, c$. Therefore,

$$(4.7) \quad \begin{aligned} \bar{A}^n(t) - \lambda t &= \bar{E}^n(t) - \alpha t + \sum_{k=1}^c \left(\bar{\Phi}^{k,n}(\bar{D}_k^n(t)) - P'_k \bar{D}_k^n(t) \right) \\ &\quad + P'(\bar{D}^n(t) - \Lambda C' \bar{\tau}^n(t)) - P' \Lambda C'(te - \bar{\tau}^n(t)), \end{aligned}$$

where we have used the fact that

$$\lambda = \alpha + P'\lambda,$$

and P_k denotes the k^{th} row of P . Using equation (4.5), we can replace $\bar{D}^n(t)$ in the third expression on the right hand side of (4.7) by $\bar{A}^n(\bar{\tau}^n(t))$ when n is large. Hence, by Lemmas 4.1 and 4.3 and functional strong law of large numbers, we have

$$\limsup_{n \rightarrow \infty} \|\bar{A}^n(\cdot) - \lambda \cdot\|_t \leq \limsup_{n \rightarrow \infty} P' \|\bar{A}^n(\bar{\tau}^n(\cdot)) - \lambda \bar{\tau}^n(\cdot)\|_t \leq P' \limsup_{n \rightarrow \infty} \|\bar{A}^n(\cdot) - \lambda \cdot\|_t.$$

Because $(I - P')^{-1} \geq 0$,

$$\limsup_{n \rightarrow \infty} \|\bar{A}^n(\cdot) - \lambda \cdot\|_t \leq 0,$$

and hence

$$\lim_{n \rightarrow \infty} \|\bar{A}^n(\cdot) - \lambda \cdot\|_t = 0.$$

\square

Lemma 4.6 *Define for each $t \geq 0$,*

$$\eta(t) = Q \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) - P' \Lambda C' W^*(t) \right).$$

Suppose the convergence in (3.7) holds. Then

$$\bar{A}^n(t) \rightarrow \eta(t) \quad \text{u.o.c. as } n \rightarrow \infty.$$

Proof. First, note that

$$\tilde{A}^n(t) = \tilde{E}^n(t) + \sum_{k=1}^c \tilde{\Phi}^{k,n} \left(\bar{A}_k^n(\bar{\tau}_{s(k)}^n(t)) \right) + P' \tilde{A}^n(\bar{\tau}^n(t)) - P' \Lambda^n C' \bar{\tau}^n(t)$$

and

$$\eta(t) = \xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) + P' \eta(t) - P' \Lambda C' W^*(t).$$

Thus,

$$\begin{aligned} \tilde{A}^n(t) - \eta(t) &= \tilde{E}^n(t) - \xi^a(t) + \sum_{k=1}^c \left(\tilde{\Phi}^{k,n} \left(\bar{A}_k^n(\bar{\tau}_{s(k)}^n(t)) \right) - \xi^k(\lambda_k t) \right) \\ &\quad + P' \left(\tilde{A}^n(\bar{\tau}^n(t)) - \eta(\bar{\tau}^n(t)) \right) + P' (\eta(\bar{\tau}^n(t)) - \eta(t)) \\ &\quad - P' (\Lambda^n C' \bar{\tau}^n(t) - \Lambda C' W^*(t)). \end{aligned}$$

Hence

$$\begin{aligned} (4.8) \quad \|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t &\leq \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t + \sum_{k=1}^c \|\tilde{\Phi}^{k,n} \left(\bar{A}_k^n(\bar{\tau}_{s(k)}^n(\cdot)) \right) - \xi^k(\lambda_k \cdot)\|_t \\ &\quad + P' \|\tilde{A}^n(\bar{\tau}^n(\cdot)) - \eta(\bar{\tau}^n(\cdot))\|_t + P' \|\eta(\bar{\tau}^n(\cdot)) - \eta(\cdot)\|_t \\ &\quad + P' \|\Lambda^n C' \bar{\tau}^n(\cdot) - \Lambda C' W^*(\cdot)\|_t. \end{aligned}$$

Because $\bar{\tau}_j^n(s) \leq s$ for all $s \geq 0$ and $j = 1, \dots, d$, we have

$$\|\tilde{A}^n(\bar{\tau}^n(\cdot)) - \eta(\bar{\tau}^n(\cdot))\|_t \leq \|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t.$$

Therefore, it follows from (4.8) that

$$\begin{aligned} (4.9) \quad (I - P') \|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t &\leq \|\tilde{E}^n(\cdot) - \xi^a(\cdot)\|_t + \sum_{k=1}^c \|\tilde{\Phi}^{k,n} \left(\bar{A}_k^n(\bar{\tau}_{s(k)}^n(\cdot)) \right) - \xi^k(\lambda_k t)\|_t \\ &\quad + P' \|\eta(\bar{\tau}^n(\cdot)) - \eta(\cdot)\|_t + P' \|\Lambda^n C' \bar{\tau}^n(\cdot) - \Lambda C' W^*(\cdot)\|_t \\ &\equiv \zeta^n(t). \end{aligned}$$

Again, note that $Q \geq 0$. Premultiplying both sides of (4.9) by Q , we have

$$\|\tilde{A}^n(\cdot) - \eta(\cdot)\|_t \leq Q \zeta^n(t).$$

Because $\zeta^n(t) \rightarrow 0$, we have proved Lemma 4.6. □

Proof of Theorem 4.1. To prove Theorem 4.1, observe that from (2.4) and (2.5),

$$\tilde{W}^n(t) = C \tilde{V}^n(\bar{A}^n(t)) + C M^n \bar{A}^n(t) + \sqrt{n}(\rho^n - e)t + \tilde{Y}^n(t).$$

By Lemmas 4.5 and 4.6 and assumptions (3.5) and (3.2), we have

$$\begin{aligned} &C \tilde{V}^n(\bar{A}^n(t)) + C M^n \bar{A}^n(t) + \sqrt{n}(\rho^n - e)t \\ &\quad \rightarrow C \xi^s(\lambda t) + C M Q \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) - P' \Lambda C' W^*(t) \right) + \beta t, \end{aligned}$$

u.o.c. as $n \rightarrow \infty$. Because the mapping defined in (2.6) is continuous, Theorem 4.1 follows immediately from the continuous mapping theorem. \square

Remark. Because a Brownian motion is almost surely not a process with bounded variation, the matrix $I + G$ must be nonsingular in order for a solution of the system (4.1)–(4.4) to exist. Multiplying both sides of (4.1) by $R \equiv (I + G)^{-1}$, one has

$$W^*(t) = RC\xi^s(\lambda t) + RCMQ \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) \right) + R\beta t + RY^*(t).$$

For each $t \geq 0$, set

$$X^*(t) = RC\xi^s(\lambda t) + RCMQ \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) \right) + R\beta t.$$

Then X^* is a Brownian motion with drift vector $\theta \equiv R\beta$ and covariance matrix

$$\Gamma^* = RC \left[\Gamma^s \Lambda + MQ \left(\Gamma^a + \sum_{k=1}^c \lambda_k \Gamma^k \right) Q' M' \right] C' R'.$$

Without worrying about the adaptness of these processes, one recognizes that W^* is a semi-martingale reflecting Brownian motion (SRBM) starting from zero defined by

$$(4.10) \quad W^*(t) = X^*(t) + RY^*(t), \quad t \geq 0,$$

and (4.2)–(4.4) with covariance matrix Γ^* , drift vector θ and reflection matrix R ; see Reiman and Williams [28] for the definition of an SRBM. Reiman and Williams [28] proved that if an SRBM exists, the matrix R must be completely- \mathcal{S} . In particular, the diagonal elements of R are positive. Conversely, Taylor and Williams [29] proved that if R is a completely- \mathcal{S} matrix, then the corresponding SRBM exists and is unique in law.

5 Proof of Theorem 3.1

Now we present the example in Dai and Wang [11] to show the limiting process W^* in Theorem 4.1 does not exist for certain networks.

Consider the two-station network pictured in Figure 1. Customers arrive at station 1 according to a Poisson process with rate α_1^n . (The index n indicates the n^{th} system.) Each customer makes 5 stops before departing from the network, and the stations visited are in the following order: 1, 1, 2, 2, 1. As explained in Section 1, we designate those customers in their k^{th} stop as class k customers. The service times for class k customers are assumed to be exponentially distributed with mean m_k ($k = 1, \dots, 5$), independent of n .

Choose $m = (1/10, 1/10, 22/27, 5/27, 8/10)'$ and

$$\alpha_1^n = \left(1 - \frac{1}{\sqrt{n}} \right).$$

Then $\alpha_1^n \rightarrow 1$, $\rho^n < e$ for each n and

$$\lim_{n \rightarrow \infty} \sqrt{n}(\rho^n - e) = (-1, -1)'.$$

For the specific data above, one can check that $\det(I+G) = 0$, thus $I+G$ is singular. Therefore, there exists a vector $u \neq 0$ such that $u'(I+G) = 0$.

Now, assume that Conjecture (3.7) is true. By Theorem 4.1, the normalized workload process and idleness process $(\tilde{W}^n, \tilde{Y}^n)$ converges to the limiting processes (W^*, Y^*) u.o.c., where

$$(5.1) \quad (I+G)W^*(t) = \xi(t) + \beta t + Y^*(t),$$

and

$$\xi(t) = C\xi^s(\lambda t) + CMQ \left(\xi^a(t) + \sum_{k=1}^c \xi^k(\lambda_k t) \right)$$

is a Brownian motion with zero drift and covariance matrix

$$\Gamma \equiv C\Gamma^s\Lambda C' + CMQ(\Gamma^a + \sum_{k=1}^c \lambda_k \Gamma^k)Q'MC'.$$

Multiplying both sides of (5.1) by u' , we get

$$(5.2) \quad u'\xi(t) = -u'\beta t - u'Y(t), \quad \text{for all } t \geq 0.$$

It is easy to check that Γ is a positive definite matrix and hence $u'\xi$ is a zero drift Brownian motion with variance $u'\Gamma u > 0$. Note that the left hand side of (5.2) is a process of bounded variation, while a Brownian motion is almost surely not a process of bounded variation. Therefore the conjecture cannot possibly hold, and Theorem 3.1 is proved. \square

Remark. If one takes

$$m = (1/10, 1/10, 23/27, 4/27, 8/10)',$$

the reflection matrix $R = (I+G)^{-1}$ becomes

$$R = \begin{pmatrix} -310/27 & 16 \\ 20 & -27 \end{pmatrix}.$$

Because the diagonal elements of R are negative, R is not a completely- \mathcal{S} matrix. Hence by Reiman and Williams [28] there is no SRBM W^* associated with the corresponding reflection matrix R . Therefore, \tilde{W}^n can not converge to an SRBM in this case. \square

Proof of Corollary 3.1 For $x \in D_{\mathbb{R}^d}[0, \infty)$, define

$$J(x) = \int_0^\infty e^{-u} [J(x, u) \wedge 1] du,$$

where

$$J(x, u) = \sup_{0 \leq t \leq u} \sum_{i=1}^d |x_i(t) - x_i(t-)|.$$

It follows from Lemma 4.2 that

$$J(\tilde{W}^n) \rightarrow 0.$$

almost surely as $n \rightarrow \infty$. Therefore, by Theorem 3.10.2 of Ethier and Kurtz [12], any limit $W^* = \{W^*(t), t \geq 0\}$ of a convergent subsequence of $\{\tilde{W}^n(\cdot), n \geq 1\}$ (under the Skorohod topology) is continuous. From the proof of Theorem 3.1, we know that such a process W^* does not exist. Therefore, $\{\tilde{W}^n(\cdot), n \geq 1\}$ cannot be D -tight. \square

6 Concluding Remarks and Open Problems

In this paper, we have proved that conventional heavy traffic limit theorems do not hold for general multiclass open queueing networks. To identify a maximal subset of multiclass networks such that the corresponding heavy traffic limit theorems prevail seems to be a formidable task for the moment. We conjecture that when there is a single service time distribution associated with each server, the convergence in (3.7) holds.

In his example, Whitt [30] demonstrated that the non-convergence of the normalized workload process may be caused by large fluctuations of the workload. In [30], these large fluctuations occur because batches of customers with short service times build up in the queues. One way to avoid such fluctuation is to employ some kind of processor sharing discipline (like head-of-the-line processor-sharing) among the customer classes at each station. It is worthwhile to investigate the heavy traffic behavior for multiclass queueing networks under non-FIFO queueing disciplines. Research in this direction is just beginning, see the appendix to Harrison and Williams [19].

Acknowledgment. We are indebted to Ruth Williams for commenting on an earlier version of this paper.

References

- [1] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260 (1975).
- [2] Billingsley, P. *Convergence of Probability Measures*. Wiley, New York, 1968.
- [3] Chen, H. and Mandelbaum, A. Stochastic discrete flow networks: Diffusion approximation and bottlenecks. *Annals of Probability* **19**, 1463–1519 (1991).
- [4] Chen, H. and Mandelbaum, A. Hierarchical modelling of stochastic networks II: strong approximations. Preprint, 1992.
- [5] Chen, H. and Shanthikumar, J. G. Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. submitted for publication.
- [6] Dai, J. G. and Harrison, J. M. Steady-state analysis of RBM in a rectangle: numerical methods and a queueing application. *Annals of Applied Probability* **1**, 16–35 (1991).
- [7] Dai, J. G. and Harrison, J. M. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Annals of Applied Probability* **2**, 65–86 (1992).
- [8] Dai, J. G. and Kurtz, T. G. Characterization of the stationary distribution for a semi-martingale reflecting Brownian motion in a convex polyhedron. Preprint.
- [9] Dai, J. G. and Kurtz, T. G. A new proof of the heavy traffic limit theorem for a multi-type one station queue with feedback. Preprint.

- [10] Dai, J. G., Nguyen, V., and Reiman, M. I. Sequential bottleneck decomposition: An approximation method for open queueing networks. submitted.
- [11] Dai, J. G. and Wang, Y. Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems: Theory and Applications* (to appear in the special issue on networks).
- [12] Ethier, S. N. and Kurtz, T. G. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.
- [13] Glynn, P. W. and Whitt, W. Departures from many queues in series. *Annals of Applied Probability* **1**, 546–572 (1991).
- [14] Harrison, J. M. Brownian models of queueing networks with heterogeneous customer populations. *Proceedings of the IMA Workshop on Stochastic Differential Systems* (1988). Springer-Verlag.
- [15] Harrison, J. M. and Nguyen, V. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications* **6**, 1–32 (1990).
- [16] Harrison, J. M. and Nguyen, V. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* (to appear in the special issue on networks).
- [17] Harrison, J. M. and Reiman, M. I. Reflected Brownian motion on an orthant. *Annals of Probability* **9**, 302–308 (1981).
- [18] Harrison, J. M. and Williams, R. J. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22**, 77–115 (1987).
- [19] Harrison, J. M. and Williams, R. J. Brownian models of feedforward queueing networks: Quasireversibility and product form solutions. *Annals of Applied Probability* **2**, 263–293 (1992).
- [20] Iglehart, D. L. and Whitt, W. Multiple channel queues in heavy traffic I. *Adv. Appl. Probab.* **2**, 150–177 (1970).
- [21] Iglehart, D. L. and Whitt, W. Multiple channel queues in heavy traffic II. *Adv. Appl. Probab.* **2**, 355–364 (1970).
- [22] Jackson, J. R. Networks of waiting lines. *Operations Research* **5**, 518–521 (1957).
- [23] Johnson, D. P. *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks*. PhD thesis, University of Wisconsin, 1983.
- [24] Kelly, F. P. *Reversibility and Stochastic Networks*. Wiley, 1979.
- [25] Peterson, W. P. A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research* **16**, 90–118 (1991).

- [26] Reiman, M. I. Open queueing networks in heavy traffic. *Mathematics of Operations Research* **9**, 441–458 (1984).
- [27] Reiman, M. I. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability* **20**, 179–207 (1988).
- [28] Reiman, M. I. and Williams, R. J. A boundary property of semimartingale reflecting Brownian motions. *Probability Theory and Related Fields* **77**, 87–97 (1988 and **80**. 633, 1989).
- [29] Taylor, L. M. and Williams, R. J. Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. Submitted.
- [30] Whitt, W. Large fluctuations in a deterministic multiclass network of queues. *Management Sciences* (to appear).

School of Mathematics and
 Industrial/Systems Engineering
 Georgia Institute of Technology
 Atlanta, Georgia 30332-0205
 USA
email: dai@isye.gatech.edu

Sloan School of Management
 Massachusetts Institute of Technology
 Cambridge, Massachusetts 02139
 USA
email: vien@mit.edu

MAR
Date Due 2001

Lib-26-67

MIT LIBRARIES

DUPL



3 9080 02239 0238

